Tips, Dones and ToDos: Uncovering User Profiles in FourSquare

Marisa Vasconcelos
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
marisav@dcc.ufmg.br

Saulo Ricci
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
saulomrr@dcc.ufmg.br

Jussara Almeida
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
jussara@dcc.ufmg.br

Fabrício Benevenuto
Universidade Federal de Ouro Preto
Ouro Preto, Brazil
benevenuto@gmail.com

Virgílio Almeida
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
virgilio@dcc.ufmg.br

ABSTRACT

Online Location Based Social Networks (LBSNs), which combine social network features with geographic information sharing, are becoming increasingly popular. One such application is Foursquare, which doubled its user population in less than six months. Among other features, Foursquare allows users to leave tips (i.e., reviews or recommendations) at specific venues as well as to give feedback on previously posted tips by adding them to their to-do lists or marking them as done. In this paper, we analyze how Foursquare users exploit these three features – tips, dones and to-dos — uncovering different behavior profiles. Our study reveals the existence of very active and influential users, some of which are famous businesses and brands, that seem engaged in posting tips at a large variety of venues while also receiving a great amount of user feedback on them. We also provide evidence of spamming, showing the existence of users that post tips whose contents are unrelated to the nature or domain of the venue where the tips were left.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services; J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms

Measurement, Human factors

Keywords

user behavior characterization, location based social networks, spamming

WSDM'12, February 8–12, 2012, Seattle, Washington, USA.
Copyright 2012 ACM 978-1-4503-0747-5/12/02 ...$10.00.

1. INTRODUCTION

Online Location-Based Social Networks (LBSNs) is a new paradigm of online social networks that has been experiencing increasing popularity, attracting millions of users. In LBSNs, users can *check in*, broadcasting their location to their friends, through the social graph. Check ins are performed in special locations, named venues, which represent physical places, such as universities, monuments, or even business enterprises and commercial brands. Check ins may be converted into points that allow users to earn badges, venue mayorships as well as receive special offers. Examples of currently popular LBSNs are Foursquare, Gowalla and Brightkite[1], out of which Foursquare is perhaps the one with the largest user population. Indeed, it has doubled the number of users in only six months, having reportedly reached the mark of 10 million registered users [25].

In Foursquare, in particular, users may also post tips to specific venues, aiming at sharing information on any aspect related to the venue with others. For instance, a tip left at a restaurant may suggest a special dish or even complain about the service. Tips can thus be seen as reviews or *recommendations*, either positive or negative, about the tipped venues. A user who sees a tip may add it to her to-do list or mark it as done, as a sign of agreement as well as a feedback regarding the tip.

As such, tips, dones and to-do lists are valuable features for fostering interactions among users, who can share their experiences and learn from others. Similarly, business owners can also benefit from tips about their stores or products, as these tips are a means for them to reach and get feedback from potential customers. Some businesses, which are Foursquare users themselves, can also use tips for promoting their brands and products [5]. Indeed, according to [26], around two thirds of the users post tips to Foursquare venues, showing the wide usage of the feature. Unlike user check ins, which propagate through the social links, and thus are visible only to the user's friends, tips are visible to everybody. Thus, tips have the potential to significantly impact online information sharing and business marketing.

A few recent studies have analyzed the properties and user

[1] http://www.foursquare.com, http://gowalla.com, and http://brightkite.com/, respectively.

characteristics of LBSNs, particularly Foursquare. However, most of these studies focus on the user check in dynamics, on the properties of the social graph and related geographical information [17, 22, 18]. We are not aware of any previous analysis of how users exploit the tip, done and to-do features. An investigation of relevant user behavior patterns, when it comes to posting tips at venues and marking them as done or to-do, is key for future designs and developments. Given that LBSNs, and Foursquare in particular, are being increasingly used as forums for recommending places, services and products as well as for Internet marketing, understanding how these features are used can produce useful insights into the potential effectiveness and vulnerabilities of these actions.

To shed some light into the current use of tips, dones and to-dos, we here present a characterization of user behavior in Foursquare. Our analyses are performed over a dataset containing over 1.6 million venues and associated information crawled from Foursquare during 8 weeks of operation (May to July 2011). For each collected venue, we gathered its tips, the users who posted each tip, the number of dones and to-dos associated with each tip as well as the Foursquare category to which the venue was assigned and its geographical location.

Our study has two main phases. First, we characterized venues and users with respect to number of tips, number of dones and to-dos as well as percentage of tips containing links (i.e., URLs or email addresses). Next, we applied a clustering algorithm to group users into profiles based on three attributes, namely, number of venues tipped by the user, total number of dones and to-dos associated with the user's tips, and the percentage of the user's tips containing links. We further analyzed the clustering results by manually inspecting the contents of the tips posted by samples of users randomly selected from each cluster.

Our study revealed four different user profiles. Two profiles correspond to users with different levels of activity in the system: whereas one corresponds to occasional users that post tips to only a few venues and receive only a few dones and to-dos, the other, containing the vast majority of the clustered users, consists of more active users who also tend to get much more feedback regarding their tips. A third profile consists of users who are characterized by tipping a large number of different venues and receiving a very large number of dones and to-dos in return. Based on the feedback received on their tips, these users, many of which are famous businesses and brands, can be considered very influential in the system.

The last profile is characterized by a very large percentage of tips with links posted at many venues, a behavior that is consistent with spamming, according to the Foursquare's terms of service [8]. Indeed, our manual inspection revealed that the majority of the tips posted by *all* sampled users with this profile had contents that were unrelated to the nature or domain of the tipped venues. Moreover, we also found users who posted unrelated tips within the other three groups, in spite of the fact that most of their tips did not contain links. This is further evidence of spamming activity. Interestingly, some of these users received a large number of dones and to-dos in their tips, in spite of them being often posted at unrelated venues. In other words, such potential spammers did get a great amount of feedback from other users, indicating that dealing with tip spamming and spammers in Foursquare is a hard task that is subject to a high degree of controversy.

The rest of this paper is organized as follows. Section 2 discusses related work, whereas Section 3 introduces the main elements and features of Foursquare, our crawling methodology and a summary of the collected dataset. Section 4 analyzes how users exploit tips, dones and to-dos, characterizing attributes of venues and of individual users, whereas the identified user profiles are presented and discussed in Section 5. Section 6 summarizes the paper and discusses possible directions for future work.

2. RELATED WORK

Social interactions among individuals located within a short physical proximity has been used to explain a number of phenomena in society, such as the proliferation of specific industries in a certain region [24] and individuals employment status [29]. In the context of online social networks, Liben-Nowell *et al.* [16] showed a strong correlation between friendship links and geographic location of those friends for LiveJournal[2] users. More recently, some articles demonstrated that geographically identified social content, like chatter from Twitter[3], can be used to monitor real-world events and create interesting applications. Particularly, Gomide *et al.* [10] proposed a spatio-temporal approach to identify potential dengue epidemics, whereas Sakaki *et al.* [19] proposed to treat Twitter users as sensors and use them to create a mechanism for earthquake detection of earthquakes.

However, there is still a limited set of studies about location-based social networks, possibly because they are still at an infant stage. Next, we briefly survey some of these studies.

Scellato *et al* [21] analyzed the social, geographic and geo-social properties of four social networks that provide location information about their users, namely BrightKite, Foursquare, LiveJournal and Twitter. They showed that LBSNs are characterized by short-distance spatial-clustered friendships, while in the other types of networks, such as Twitter and LiveJournal, users have heterogeneous connection lengths. An analysis of Gowalla users showed that the number of friends follows a double Pareto-like distribution, whereas the numbers of check ins and places are better described by log-normal distributions [20]. The authors also analyzed the temporal variations of such distributions, observing that users tend to add new friends at a faster rate than they give check ins and go to new places. In [17], the authors analyzed the user check in dynamics and the presence of spatio-temporal patterns in Foursquare.

The link prediction problem in LBSNs, representing the task of recommending friends or places, has also been tackled. Using collaborative filtering techniques, Berjani and Strufe [4] proposed a personalized recommender for places in Gowalla based on the number of check ins at spots. In [15], the authors proposed a three-layered friend recommender model using attributes of user profiles (preferences), social graphs (friendship) and mobility patterns (distance of visited places) to recommend friends in Brightkite. Another supervised learning framework to recommend places and friends was proposed in [23] and evaluated in a longitudinal dataset collected from Gowalla. Using check in information regarding users who visited the same place and friends of

[2] http://www.livejournal.com

[3] http://www.twitter.com

friends, the authors were able to reduce the link prediction space. Both previous studies [23, 15] concluded that the inclusion of information about location-based activity leads to a better prediction than if only social data is used.

Human mobility patterns were investigated on Gowalla, Brightkite and cell phones trace datasets [6]. The authors observed, in the three datasets, that short-range spatially and temporally periodic movement is not influenced by the social network structure while the long-distance travel is more influenced by the social links. Thus, they proposed a mobility model that combines periodic daily movement patterns and the social movement effects caused by the friendship network.

We are aware of only two previous studies that aim at uncovering user profiles in LBSNs. In [14], the authors applied two different clustering approaches to identify user behavior patterns on BrightKite. One approach exploited the users update (i.e., check ins, photos and notes) geographic position to classify them into four groups according to their mobility, namely, home, home-vacation, home-work and other. The second approach clustered users based on multiple attributes such as total number of updates, social features and mobility characteristics, and led to the identification of five groups, namely, inactive, normal, active, mobile and trial (or non loyal) users. The second study was performed on Foursquare [18]. The authors used a spectral clustering algorithm to group users based on the categories of venues they had checked in, aiming at identifying communities and characterizing the type of activity in each region of a city.

Although the aforementioned studies offer important insights into properties of user interactions in LBSNs, none of them addressed how users exploit tips, dones and to-dos as means to review or recommend a place or a service as well as to give feedback regarding previously posted tips. Our work aims at contributing to fill this gap.

3. FOURSQUARE DATASET

In this section, we first describe Foursquare, its main elements and features. In particular, we introduce the tip, done and to-do features, which are target of our study (Section 3.1). Next, in Section 3.2, we describe the strategy adopted to crawl Foursquare and summarize the collected dataset used in the following sections.

3.1 Foursquare

Foursquare is a prime example of online Location-Based Social Network (LBSN) on which users share their location with friends. As videos and images are the main objects on YouTube and Flickr, respectively, the main objects on Foursquare are venues. A **venue** represents any physical location like a store, a restaurant, an university, an airport or a monument, where users can check in. Users may check in venues when they are physically close to those venues, using GPS-equipped mobile devices or SMS messages. Once users check in, they may choose to share their location with friends. Every time a user checks in a venue, she collects points, namely **badges**, on Foursquare. If a user has more "check ins" in a certain venue than any other user, she becomes the venues' mayor. Venues are created by Foursquare users who become owners of that place. However, venues can be claimed by real business owners. In this case, venues are verified by Foursquare and, if approved, the real owners of the venue can start offering promotions and special deals to users that frequently check in that venue. Foursquare also maintains a set of eight pre-defined venue categories, namely, "Arts & Entertainment", "Colleges & Universities", "Food", "Great Outdoors", "Nightlife Spots", "Travel Spots", "Shops", "Home, Work and Others".

In addition to checking in venues, users are also allowed to post tips at specific venues for other users that visit them in the future. **Tips** are pieces of information containing *recommendations* (or reviews), either positive or negative, about the venue. Examples of tips include the best option of a menu in a restaurant, the best place to have lunch in an airport, or even a complaint about a service in a venue. Tips have become extensively used in Foursquare as they allow one to share her opinion about nearby physical locations with others. When a user posts a tip, her immediate friends receive it as a message. Users may also see previously posted tips by searching nearby locations or accessing the venues.

After reading a tip, a user may add it to her **to-do** list or even mark it as **done**. Thus, tips, dones and to-dos foster interactions among users who accessed the same venue in the system. An illustrative example of such interactions is presented in Figure 1. User Ann posts a tip at a certain venue. After a while, user John searches for nearby locations, finds the venue tipped by Ann, reads it, and decides to add it to his to-do list. John may also mark the tip as done to indicate that he followed and agreed with Ann's recommendation. Thus, the total number of times a tip was marked as done or added to a user's to-do list (here referred to as, simply, the number of dones and to-dos the tip received) can be seen as an estimate of the amount of *feedback* from other users regarding that tip. Indeed, Foursquare maintains statistics about these numbers as an attempt to allow users to identify potentially good tips to follow.

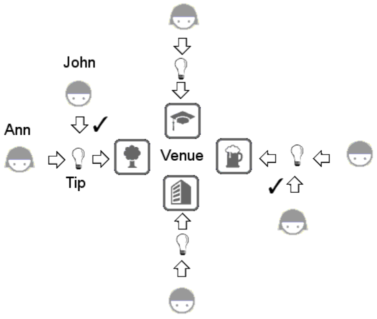


Figure 1: User Interactions Through Tips, Dones, and To-Dos: An Illustrative Example

Finally, it is important to mention that, in addition to ordinary user accounts, Foursquare also allows the creation of a special type of account, named brand user. **Brand users** are known business owners (e.g., "the History Channel", "Starbucks" and "The Wall Street Journal", among others) that are in the system to create content in the form of tips, comments, and photos [1], usually aiming at promoting themselves and their products. Towards that goal, they usually create special offers to attract users to their locations. Unlike ordinary users, who have friends, brand users have followers. The followers of a brand user receive all the tips posted by it.

3.2 Dataset

Ideally, we would like to have at our disposal data for each existing venue, including their tips, the user who posted each tip, the users who added each tip to their to-do lists, as well as the users who marked it as done. Since this data is not publicly available, we chose, instead, to crawl Foursquare to obtain a sample of it.

We used the Foursquare API to gather data about a sample of the existing venues and the users who interacted with them[4]. Our crawling strategy was based on the fact that each venue in Foursquare is assigned a unique and sequential numeric identifier (ID). Given (an estimate of) the largest ID assigned to a venue in the system, M, our crawler first randomly selects an ID between 0 and M, according to a uniform distribution. The random selection was performed to minimize the chance of bias in the sampled venues. Next, it sends a request to the Foursquare API to gather information about the corresponding venue. In particular, for each collected venue, the crawler collects all its tips, the identifications of the users who posted each of them, the number of dones and to-dos each tip received, the number of users who checked in the venue, the venue category as well as its geographic coordinates.

In order to estimate M, we ran a series of initial experiments consisting of sending HTTP GET requests to the pages of specific venues identified by their IDs. We tried increasing values of IDs, starting with 0. The largest ID for which we did get a response corresponding to a valid webpage was 20 million. We experiment with many IDs greater than that. In all those cases, the response was "Not Found". Thus, we speculate that, by the time of these experiments, this was the largest venue ID in the system, and was used as input to our crawler.

Our crawler ran from May 23^{rd} to July 19^{th}, gathering data corresponding to more than 1.6 million venues. Table 1 summarizes the collected dataset. Associated with the crawled venues, we were able to identify almost 1 million tips and more than half million unique users, out of whom 1,248 are brand users. Moreover, 3.8% of the venues are verified whereas 18.5% of them had received, by the time of the crawling, at least one tip. Since our focus is on understanding how users exploit tips, dones and to-dos, our analyses in the following sections are performed over the venues with at least one tip.

Table 1: Summary of our Foursquare Dataset

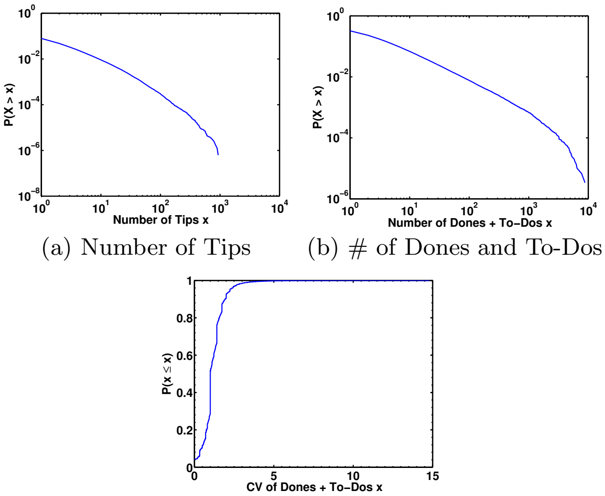
Number of venues	1,601,412
Number of venues with at least one tip	296,217
Number of verified venues	61,378
Number of users	526,651
Number of brand users	1,248
Number of tips	984,251
Total number of dones for all tips	1,407,835
Total number of to-dos for all tips	393,574

4. USER INTERACTIONS THROUGH TIPS, DONES AND TO-DOS

In this section, we analyze how users exploit the tip, done and to-do features in Foursquare. We start by focusing on the venues, and analyze how users interact with them by posting tips and by marking them as done or to-do (Section 4.1). Next, we discuss how users interact through tips and how their tips are evaluated (Section 4.2).

4.1 Venue Analyses

We here characterize each venue in our dataset in terms of the number of tips and the total number of dones and to-dos associated with all the tips posted at the venue. The former represents the amount of reviews (i.e., recommendations) from the users regarding the venue, whereas the latter is here taken as an estimate of the amount of user feedback regarding those reviews. Figures 2(a) and 2(b) show the Complementary Cumulative Distribution Functions (CCDFs) of these measures. Note the log scale in both axes.



(a) Number of Tips

(b) # of Dones and To-Dos

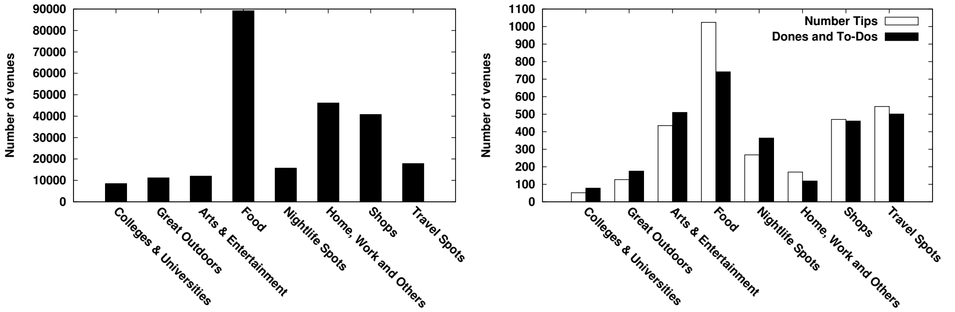
(c) CV of Number of Dones and To-Dos per Tip

Figure 2: Distributions of Venue Attributes

Figure 2(a) shows that around 57% of the venues have only one tip, whereas, some venues (around 500) are very popular among the users, receiving more than 100 tips each. Similarly, Figure 2(b) shows that there are a few venues (around 200) with tips that received a lot of feedback, with a total number of dones and to-dos exceeding 1,000. The Spearman's rank correlation coefficient, which is a non-parametric measure of statistical dependence between two variables [13], computed over the number of tips and the total number of dones and to-dos of each venue is 0.6. This reasonably strong positive correlation implies that tipping tends to be an effective mechanism to attract visibility to a venue: the larger the number of tips, the higher the amount of user feedback the venue tends to receive. Moreover, we also found a reasonably strong correlation (Spearman coefficient of 0.56) between the number of dones and to-dos and the number of distinct users who checked in each venue, indicating that user feedback is somewhat related to venue visitation in the system (i.e., check ins).

Recall that dones and to-dos are associated with tips.

[4] We note that, alternatively, one might choose to crawl Foursquare by requesting user data. We opted to collect venues and their associated data instead, because user data is not publicly available.



(a) All Venues

(b) Top 1% Venues According to Number of Tips and Number of Dones and To-Dos

Figure 3: Distribution of Venues Across Categories

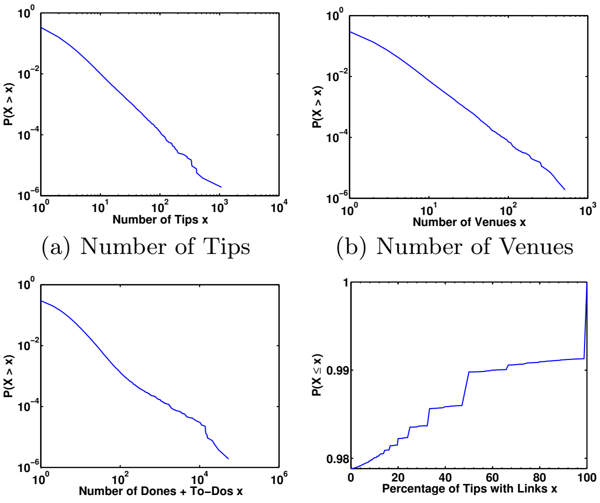
Thus, different tips posted at the same venue may attract different feedback. To investigate this issue, we analyze the variability in the number of dones and to-dos across different tips of the same venue. Figure 2(c) shows the Cumulative Distribution Function (CDF) of the coefficient of variation - CV (ratio of standard deviation to the mean) of this number. We considered only venues with at least 2 tips and at least one done/to-do. Note that most (70%) of the venues have CVs larger than 1, and around 9% of our venues have CVs larger than 2, indicating reasonably high variability. These numbers imply that user feedback regarding a given venue does depend on the specific tip, and some tips do tend to attract more feedback than others.

As discussed in Section 3.1, Foursquare venues are grouped into 8 pre-defined categories. Figure 3(a) shows a histogram of the number of venues in each category, in our dataset. To understand which venue category attracts the largest number of tips, dones and to-dos, we ranked the venues according to the number of tips and according to the total number of dones and to-dos. We then selected the venues in the top 1% of each rank. Figure 3(b) shows the number of venues of each category in each selection. Note that "Food" contains the largest number of venues both in the top 1% more tipped venues (1024) and in the top 1% venues with more dones and to-dos (742). The second category with more venues is "Travel Spots" in the first rank and "Arts & Entertainment" in the second one. Indeed, the two venues with largest number of tips are "Super Bowl Event" ("Arts & Entertainment") and "Jakarta Airport" ("Travel Spots"), whereas "Grand Central Terminal" ("Travel Spots") and "Madison Square Garden" ("Arts & Entertainment") received the largest numbers of dones and to-dos.

4.2 User Analyses

We now turn our focus to the users, analyzing how they exploit tips, dones and to-dos. In particular, we characterize the number of tips posted by each user, the total number of dones and to-dos received by her tips, the number of venues tipped by her and the percentage of her tips containing links (i.e., URLs and e-mail addresses). The distributions of these measures are shown in Figure 4. Note that Figures 4(a,b,c) show CCDFs with both axes in logarithm scale.

Figures 4(a-b) shows that, like for the number of tips and the number of dones and to-dos per venue (Figures 2(a-b)), the the distributions of number of tips and number of tipped venues per user are heavy tailed. Most users posted a few tips at a few venues, whereas a few, very active users posted many tips at many different venues. Indeed, 66% of the users posted only one tip and 70% of the users posted tips at only one venue[5], whereas 67 users posted more than 100 tips and 39 posted tips at more than 100 venues. The distribution of the total number of dones and to-dos per user (Figure 4-c) is also heavy-tailed: around 103,348 users received only one to-do/done in their posted tips, whereas 90 users received at least 1,000 dones and to-dos.



(a) Number of Tips

(b) Number of Venues

(c) # of Dones and To-Dos

(d) % of Tips with Links

Figure 4: Distributions of User Attributes

The distribution of the percentage of tips containing links, shown in Figure 4(d), has a different pattern: the vast majority (98.7%) of all users who posted at least one tip did not add any link in their tips' contents. However, we do notice the presence of some users (around 200) with more than 60% of their tips containing links. These users are further discussed in Section 4.2.2.

We also analyzed how the four user attributes correlate. We found no significant correlation between the percentage of tips with links and any other attribute. In contrast, we did find a strong positive correlation (Spearman coeffi-

[5] Note that, since we collected users through venue tips, all analyzed users have posted at least one tip.

cient of 0.93) between the number of tips and the number of venues tipped by a user, suggesting that users who more often add tips do tend to spread them across more venues. The correlation remains very high (above 0.78) even when computed over selected groups of very active users, i.e., over the top 1% users with the largest number of (1) tips, (2) tipped venues, and (3) dones and to-dos in their tips.

We also found that the correlation between the number of tips and the total number of dones and to-dos received by a user's tips is only moderate (0.40), being even lower (0.23-0.30) if computed over the three aforementioned subsets of very active users. Thus, in general and in contrast to our finding regarding venues, users who tip more do not necessarily receive more feedback. If we take the total number of dones and to-dos received by a user's tips as an estimate of her influence in the system, such influence is only moderately correlated with her degree of tipping activity, seeming much more related to the tipped venue. We further analyze user influence next.

4.2.1 User Influence

Figure 5(a) shows the number of tipped venues versus the total number of dones and to-dos received by each user. Note the logarithm scale in both axes. In this plot, users are grouped into two sets based on the maximum geographical distance between any pair of venues tipped by them. We refer to such distance as the *diameter* of the venues tipped by the user, and use it to assess the scale (local or global) of the user's influence. We group users into those with diameter shorter than or larger than 40 kilometers. Figure 5(b) shows a similar graph, plotting the number of tips versus the total number of dones and to-dos of each user. To improve graph readability, both figures only show users with at least 10 tips. Note that, given the strong correlation between number of tips and number of tipped venues per user, the two graphs are similar.

Both graphs show the existence of different classes of users. On one hand, we find users who, despite posting a large number of tips and/or posting tips to a large number of venues, receive only a comparatively small number of dones and to-dos. One such user, marked as "User 1" in both graphs, posted approximately 104 tips at 100 different venues, but received only 4 to-dos and dones. By manually inspecting a sample of those users, located in the right bottom corner of the graphs, we found that they tend to be ordinary users who, in spite of being very active in the system and contributing with a lot of tips regarding many different venues, do not get a lot of feedback from others.

In contrast, the right top corner of the plots show users who not only post a large number of tips at a large number of different venues but also receive a lot of feedback on them. Those users, many of which are famous brands such as "Bravo", "History Channel", "The Wall Street Journal" and "National Post", seem engaged in providing many recommendations (tips) on a large variety of venues, and clearly succeed in reaching and attracting the attention of many users. Clearly, those users are very influential in the system. Note that, interestingly, the graphs show users with large number of tips (and tipped venues) and large number of dones and to-dos in both user sets, suggesting the presence of local and global influence.

Both graphs also show some users who, despite posting only a few tips at a few venues, received a comparatively very large amount of feedback, and thus can also be considered very influential. Examples are "User 2" and "User 3" in Figure 5(a), who received 2708 and 1337 dones and to-dos, respectively, despite targeting only a couple of venues with a few tips. Some of these highly focused influential users are brands, such as "Six Flags", while others are ordinary users (e.g., "User 4"). Once again, we found focused users with strong influence both locally and globally.

As a side note, we point out that, as expected, both graphs show a slight trend for users with larger numbers of tips and tipped venues also having larger diameter. Perhaps more interesting is the presence of some very active users, with tens to hundreds of tips and tipped venues, who are focused on venues in a local region (diameter under 40 kilometers).

4.2.2 Suspicious Behavior

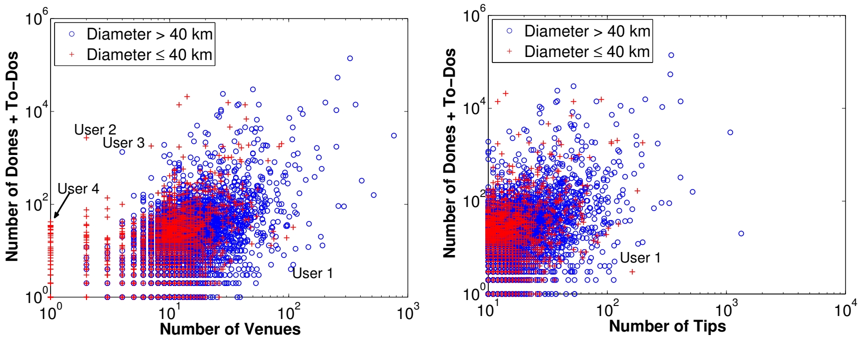
According to the Foursquare's Terms of Service, the introduction of links to unrelated sites across various venues is considered *spamming*, and users who are caught doing it will have their accounts deactivated[8]. As previously discussed, we did find some users who had the majority of their tips containing links, i.e., URLs and email addresses. This finding raised a concern about suspicious behavior, particularly because only links included in the tip's text were accounted for. In other words, links placed in the "More info" field, expected to be directly related to the target venue (e.g., the venue's website), as well as related pictures, placed in a separate field, were disregarded.

To delve deeper into this issue, we selected users with at least 10 tips and at least 60% of their tips containing links for further analysis. This selection corresponded to 3% of all users that posted at least 10 tips. Figure 6(a) plots the percentage of tips with links versus the number of tipped venues for these users, whereas Figure 6(b) shows the percentage of tips with links versus the total number of dones and to-dos. As in the previous section, we group users based on the diameter of their tipped venues.

The graphs show no clear correlation between the percentage of tips with links and the total number of dones and to-dos or the number of tipped venues. Indeed, the Spearman coefficients are -0.17 and 0.13, respectively. In other words, there are many users with a large percentage of tips with links that posted tips at only a few venues, which cannot necessarily be configured as spamming according to Foursquare's rules. Moreover, there are also users that, despite the large percentage of tips with links, did receive a large number of dones and to-dos (see discussion below).

However, Figure 6(a) also shows that several of the selected users posted tips with links at a large number of venues. Take for instance "User 5", "User 6" and "User 7" in that figure. They posted tips at more than 100 different venues, and all tips contained links. These numbers reveal a behavior pattern that is consistent with spamming, and violates Foursquare's Terms of Service. Moreover, the total numbers of dones and to-dos for "User 5" and "User 7" are only 12 and 32, respectively, whereas "User 6" received no feedback. Interestingly, we found no clear correlation between suspicious behavior and diameter of the tipped venues. In other words, our results reveal potential spamming activity both locally and globally.

We note that not all users who posted tips with links at many venues are necessarily engaged in spamming activity, as the webpage linked to might be somewhat related to the



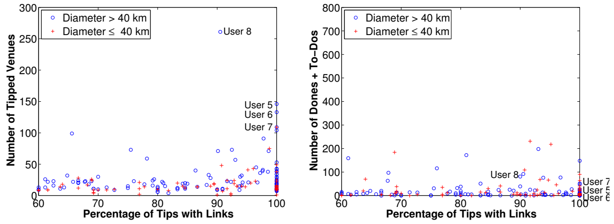
(a) Number of Venues vs. Number of Dones and To-Dos

(b) Number of Tips vs. Number of Dones and To-Dos

Figure 5: Correlation between User Attributes (only users with at least 10 tips)

tipped venue. Take, for instance, "User 8" in Figure 6(a), that posted 286 tips, 90% of which containing links, at 261 different venues. Despite the large number of tips with links, those tips received a total 92 dones and to-dos. We manually investigated this user, finding that it corresponds to a large business chain that placed the same tip to all of its stores advertising a promotion. The tip contained a link to an external webpage that should be visited by those interested in participating to learn more about it. A reasonably large number of users (92) liked it enough to mark the tip as done or add it to their to-do lists. In this case, the link pointed to a content that was related to the tipped venue.

Moreover, as we further discuss in Section 5.2, some users that post many tips containing links at many venues aiming at spamming might still receive many dones and to-dos from others. In other words, they might succeed in triggering the interest of many users.



(a) # Tipped Venues vs. % of Tips with Links

(b) # Dones and To-Dos vs. % of Tips with Links

Figure 6: Correlation between User Attributes (top 3% users with largest percentages of tips with links)

5. USER PROFILES

In the previous section, we discussed various user behavior patterns observed in our Foursquare dataset, with respect to the use of tips, dones and to-dos. We here go one step further and identify user profiles. We do so by applying a clustering algorithm to group users based on three attributes, namely the number of tipped venues, the total number of dones and to-dos and the percentage of tips with links[6].

We selected the Expectation-Maximization (EM) clustering algorithm, which is a well-known algorithm used for clustering in the context of mixture models [7].We ran the EM implementation in Weka [30], which has a built-in iterative mechanism to determine the number of clusters. The mechanism is based on ten-fold cross-validation: for each candidate number of clusters, it breaks the data into 10 folds, 9 are used as training sets and 1 as testing set. It builds the clusters on the training sets and, given those clusters, computes the log-likelihood for each instance in the testing set. The log-likelihood values are summed and then averaged over all 10 folds. The number of clusters selected is the one with maximum (average) log-likelihood.

Because of the large variability observed in the values of the user attributes, particularly number of tipped venues and total number of dones and to-dos, which make the clustering task harder, we converted all values to a log scale, and normalized the results afterwards.

Next, we first present the clustering results (Section 5.1) and then discuss some findings of a manual inspection of selected users (Section 5.2).

5.1 Clustering Results

We applied the EM clustering algorithm over *all users with at least 10 tips*. The algorithm identified 4 clusters, referred to as, throughout this section, clusters 0, 1, 2 and 3. Table 2 shows, for each cluster, averages and CVs for each user attribute. It also shows the number of users in each cluster. Complementarily, Figure 7 shows, for each cluster, the CDF of each attribute.

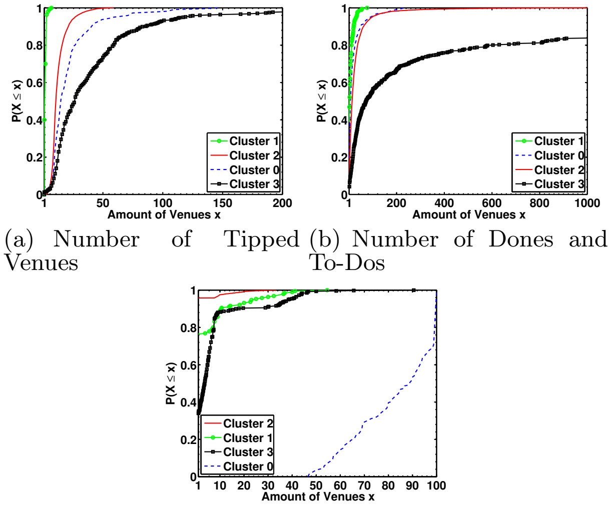
Cluster 0, which includes around 3% of all clustered users, is characterized by a much larger percentage of tips with links (83% on average). This is consistent across most users of the cluster, as shown in Figure 7(c) and summarized by the low CV. Indeed, this attribute clearly distinguishes these users from the others. The number of tipped venues also tends to be large, though smaller than for users in cluster 3. These patterns are consistent with the suspicious, potential spamming, behavior discussed in Section 4.2.2. Moreover, in general, users in cluster 0 do not tend to receive a large number of dones and to-dos.

Cluster 1 consists of focused users who are neither very active nor influential: they tend to post tips at only a few venues and do not receive many dones and to-dos from oth-

[6] We did not include the number of tips as an attribute because of the strong correlation between number of tips and number of tipped venues (see section 4.2).

Table 2: Summary of User Attributes Across Clusters

Attribute	Cluster 0		Cluster 1		Cluster 2		Cluster 3	
	avg	cv	avg	cv	avg	cv	avg	cv
Number of Venues	21.99	0.94	1.97	0.52	13.23	0.52	43.81	1.41
Percentage of Tips with Links	83.11	0.20	3.88	2.35	0.62	5.21	7.02	1.71
Number of Dones and To-Dos	20.41	1.82	7.35	1.52	29.53	2.09	1350.58	5.48
Number of Users	222		190		5660		477	



(a) Number of Tipped Venues (b) Number of Dones and To-Dos

(c) Percentage of Tips with Links

Figure 7: User Attribute Distributions

ers. These are mostly occasional users[7]. Users of cluster 2, on the contrary, are much more active: they tend to tip a larger number of venues, mostly with no links, getting many more dones and to-dos in return. Around 86% of all clustered users are in this group.

Finally, cluster 3, containing around 7% of the considered users, is characterized by the largest total number of dones and to-dos. These users also tend to post tips a large number of venues. Therefore, we expect that most very influential users that target a large number of venues fall into this cluster. Moreover, as shown in Figure 7(a), this cluster also contains users that post tips at only a few venues, indicating that this cluster also contains some very influential but focused users.

We also analyzed the distribution of venues tipped by users in each cluster across the several categories maintained by Foursquare. Figure 8 shows the distributions. The fractions of venues tipped by users from both clusters 0 and 3 vary only slightly across categories, except for "Colleges & Universities", which clearly attracts much fewer tips than the other categories. Indeed, it is the least popular category among users of all four clusters. In other words, neither users who seem engaged in suspicious activity (cluster 0) nor those who tend to be very influential in the system (cluster 3) are focused, collectively, on specific categories. In contrast, the venues tipped by the occasional users (cluster 1) are more concentrated in the "Food" category. The same can be said, to a lesser extent, for users of cluster 2.

5.2 Manual Inspection

As discussed in Section 4.2.2, we did find evidence of suspicious behavior, consistent with spamming activity. Note that this evidence was based only on the percentage of a user's tips containing links to external sites and email addresses and, to a lesser degree, on the feedback those tips received from other users. However, an interested spammer may find other ways of reaching users. For instance, a spammer interested in selling a product may write a text advertising it, and post it as a tip. One such case was a tip posted to various venues, including a japanese restaurant and an university, whose contents advertised a fitness center. The tip's text is completely unrelated to the nature and business domain of the venue.

We here consider as spam a tip whose content is unrelated to the tipped venue, typically an advertisement for a product that is, in nature, unrelated to that kind of venue. Given this definition, we note that, some spam tips might indeed be successful in getting a large number of dones and to-dos, since many users may find the advertised product interesting, despite it being unrelated to the venue where the tip was posted.

To further investigate the existence of tip spamming in Foursquare, we manually inspected a sample of users from each cluster. For each sampled user, we inspected the contents of her tips and the venues at which those tips were posted. In case the tip contained a link, the contents of the page pointed to by the link were also inspected.

Each sampled user was inspected independently by three volunteers, who labeled the user as either spammer or not. A user was labeled spammer if the contents of at least 50% of her tips were not related with the nature or domain of the tipped venues[8]. The volunteers were instructed to be conservative: in doubt, they should label the user as *not spammer*. Majority voting was used for final classification, although the volunteers agreed in the vast majority (93%) of the cases. The volunteers also counted down the number of inspected users who are brand users.

Table 3 presents our results, showing, for each cluster, the number of inspected users, the number and percentage of them labeled as spammers by the volunteers, and the number and percentage of them identified as brands. The sample size for each cluster was defined so as to have a maximum error in our estimates of 5% with 95% confidence [12].

Most of the users labeled as spammers are, as expected, in the cluster 0 sample. Indeed, all users sampled from

[7] We did observe, among users of cluster 1, many tips posted at the same venue within a very short time. For instance, around 10% of users of cluster 1 had an average inter-tip time below 2 seconds, suggesting that the user might have posted the same tip multiple times without knowing. This might reflect lack of experience with the application.

[8] Recall that, given our data collection methodology, our analysis is based only on a subset of all tips posted by each user.

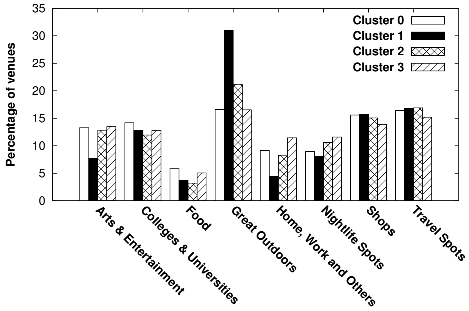


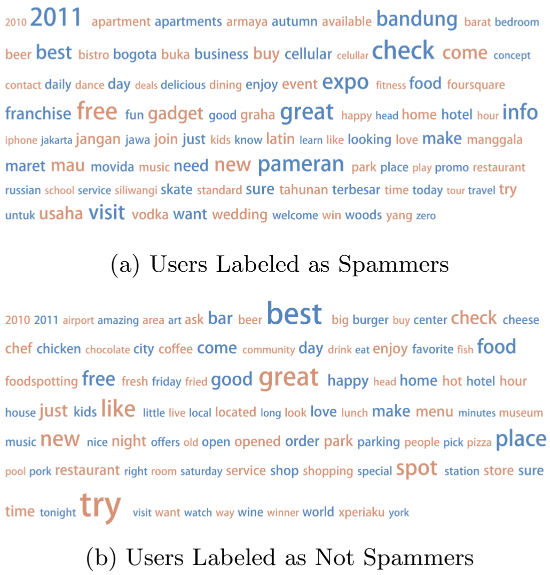
Figure 8: Venue Category Distributions

that cluster were labeled as spammers, mostly because they posted tips with links pointed to unrelated content. However, our results also show the presence of spammers in clusters 1, 2 and 3. Most of them were classified as such because the text of their tips advertised a product unrelated to the nature of the venue. Moreover, some users of cluster 3 labeled as spammers did receive a large number of dones and to-dos, indicating that, despite posting unrelated content, they did trigger the interest of many users[9].

Table 3 also shows that there are brand users in all four clusters, although the vast majority of them are in cluster 3. As discussed in Section 3, brand users are special Foursquare users that are expected to use the system to reach their followers and potential customers. The brand users of cluster 3 indeed succeeded in promoting themselves in Foursquare: they are very influential in the system, receiving a large number of dones and to-dos. Interestingly, we also found 4 brand users among the 127 users sampled from cluster 0 that were labeled as spammers.

We also analyzed the words commonly used by users labeled as spammers in their tips, contrasting them with the vocabulary of the remaining users. Figure 9 shows the word clouds of both user groups. Note that words related to service or product advertisement, such as *business*, *apartment*, *cellular*, *franchise*, *gadget* e *iphone* are more frequently used by users labeled as spammers.

Finally, we note that the concept of spamming in Foursquare is a very subjective and thus controversial matter, possibly even more than in other systems. What one classifies as spam, another person might interpret as creative marketing strategy. Take for instance the case, reported in [5], of a tip advertising a product sold by a certain venue v_i, posted at another venue v_j, which is indeed a business competitor of v_i. When customers check the tips left at v_j, they will see the advertisement and might indeed be drawn towards its competitor. In other words, the tip might contribute to steal business away from the venue at which it was left[10]. Moreover, obliviously of any marketing game, users might find the tip interesting and mark it as a to-do or done. This might happen even for tips advertising products completely unrelated to the tipped venues, as discussed above. Thus, identifying and dealing with tip spamming and spammers is a hard task. The design of automatic detection mechanisms is a challenge we intend to pursue in the future.


2010 2011 apartment apartments armaya autumn available bandung barat bedroom
beer best bistro bogota buka business buy cellular celullar check come concept
contact daily dance day deals delicious dining enjoy event expo fitness food foursquare
franchise free fun gadget good graha great happy head home hotel hour info
iphone jakarta jangan jawa join just kids know latin learn like looking love make manggala
maret mau movida music need new pameran park place play promo restaurant
russian school service siliwangi skate standard sure tahunan terbesar time today tour travel try
untuk usaha visit vodka want wedding welcome win woods yang zero
2010 2011 airport amazing area art ask bar beer best big burger buy center check cheese
chef chicken chocolate city coffee come community day drink eat enjoy favorite fish food
foodspotting free fresh friday fried good great happy head home hot hotel hour
house just kids like little live local located long look love lunch make menu minutes museum
music new nice night offers old open opened order park parking people pick pizza place
pool pork restaurant right room saturday service shop shopping special spot station store sure
time tonight try visit want watch way wine winner world xperiaku york

(a) Users Labeled as Spammers

(b) Users Labeled as Not Spammers

Figure 9: Words Commonly Used in Users' Tips

6. CONCLUSIONS

This paper analyzes how users exploit tips, dones and to-dos in Foursquare. Based on a crawled dataset containing over 1.6 million venues and more than half million users, we have identified four groups of users with very different behaviors. Two profiles correspond to regular users that differ in terms of their levels of tipping activity in the system. A third profile consists of users that seem engaged in posting tips at a large variety of venues. These users, some of which are famous businesses and brands, are typically very influential in the system, receiving a large amount of feedback from others regarding their tips. Finally, we have identified a group of users characterized by posting tips containing links at many different venues. A manual inspection of a sample of these users confirmed them as potential spammers as they post tips that are unrelated to the venue. However, we also show that some spammers do succeed in attracting the attention of many users.

Our findings unveil various interesting findings with important implications. Two such findings are: (1) we provide the first pieces of evidence of spamming activity in Foursquare. Spam has been observed in many other online social systems, including Facebook [9],YouTube [3], and Twitter [11]. As a result, a number of efforts towards designing effective strategies to detect and remove spam from these systems are available [2, 28]. Although it is debatable whether the kind of spamming activity we uncovered here corresponds to malicious/opportunistic acts that deserves punishment, we hope that our analyses serve as motivation for future discussions on the matter and guide future developments to proper address such actions on Foursquare.

Finally, it is known that the democratization of technologies like Foursquare is fundamentally changing the way people interact with each other as well as with local opinion

[9] We note that most of these users received only a couple of dones and to-dos in each tip. However, because they posted a large number of tips at various venues, collectively, those tips end up attracting a large number of users.

[10] During our manual inspection, we took a conservative approach and *did not* consider such cases spamming.

Table 3: Results from Manually Inspecting a Sample of Users from Each Cluster

Cluster	Number of Sampled Users	Number of Spammers	Number of Brands
0	127	127 (100%)	4 (3.2%)
1	181	13 (7.2%)	2 (1.1%)
2	237	37 (15.6%)	7 (2.9%)
3	203	41 (20.2%)	58 (28.6%)

leaders, small businesses, and online customers. Unlike other social networks, Foursquare virtual interactions may reverberate in the real world. When a user posts a tip, he is interacting with a real business and exposing weaknesses or qualities about this business to everyone, including competitors. Thus, tips, dones and todos are valuable tools for registered businesses as they are sources of candid feedback from their customers, provide opportunity for improvements and may greatly impact future visitations and ultimately revenues. For the users, tips may act as filters to help them choose places to visit. While there are many skeptics who doubt that social networks will generate profits that match expectations [27], our study provides early pieces of evidence of the true "social" opportunity that lies on Foursquare in the role of brands, venues, local influential users and even opportunistic users that act like spammers. Furthermore, by analyzing the roles that different types of users play in Foursquare, we hope our findings may guide the design of future recommendation approaches.

7. ACKNOWLEDGMENTS

This research is partially funded by the Brazilian National Institute of Science and Technology for the Web (MCT/CNPq/INCT grant number 573871/2008-6), and by the authors' individual grants from CNPq, CAPES and FAPEMIG.

8. REFERENCES

[1] 360i. Foursquare Brand Pages. http://blog.360i.com/social-media/foursquare-brand-pages.

[2] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Goncalves. Detecting Spammers and Content Promoters in Online Video Social Networks. In *Proc. ACM SIGIR'09*, 2009.

[3] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross. Video Interactions in Online Video Social Networks. *ACM TOMCCAP'09*, 5(4):1–25, 2009.

[4] B. Berjani and T. Strufe. A Recommendation System for Spots in Location-Based Online Social Networks. In *Proc. of SNS'11*, 2011.

[5] Catalyst Marketers Blog. Catalyst Marketers Blog. http://www.catalystmarketers.com/.

[6] E. Cho, S. Myers, and J. Leskovec. Friendship and Mobility: User Movement In Location-Based Social Networks. In *Proc. of KDD'11*, 2011.

[7] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[8] Foursquare. Foursquare house rules. http://support.foursquare.com/entries/386768-foursquare-house-rules.

[9] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and Characterizing Social Spam Campaigns. In *Proc. ACM IMC'10*, 2010.

[10] J. Gomide, A. Veloso, W. M. Jr., V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira. Dengue Surveillance Based on a Computational Model of Spatio-Temporal Locality of Twitter. In *Proc. of the ACM WebSci'11*, 2011.

[11] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @Spam: the Underground on 140 Characters or Less. In *ACM CCS'10*, 2010.

[12] R. Jain. *The Art of Computer Systems Performance Analysis: techniques for experimental design, measurement, simulation, and modeling.* Wiley, 1991.

[13] M. Kendall and J. D. Gibbons. *Rank Correlation Methods.* A Charles Griffin Title, fifth edition, 1990.

[14] N. Li and G. Chen. Analysis of a Location-based Network. In *Proc. of IEEE CSE'09*, 2009.

[15] N. Li and G. Chen. Multi-layer Friendship Modeling of Location-based Mobile Social Networks. In *Proc. of MobiQuitous'09*, 2009.

[16] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic Routing in Social Networks. In *Proc. of PNAS*, volume 102, pages 11623–11628, 2005.

[17] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proc. of ICWSM'11*, 2011.

[18] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *Proc. of SMW'11*, 2011.

[19] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. In *Proc. of WWW'2010*, 2010.

[20] S. Scellato and C. Mascolo. Measuring User Activity on an Online Location-based Social Network. In *Proc. of NetSciCom'11*, 2011.

[21] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora. Distance Matters: Geo-social Metrics for Online Social Networks. In *Proc. of WOSN'10*, 2010.

[22] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial Properties of Online Location-based Social Networks. In *Proc. of ICWSM'11*, 2011.

[23] S. Scellato, A. Noulas, and C. Mascolo. Exploiting Place Features in Link Prediction on Location-based Social Networks. In *Proc. of KDD'11*, 2011.

[24] O. Sorenson. Social Networks and Industrial Geography. *Journal of Evolutionary Economics*, 13(5):513–527, 2003.

[25] Techcrunch. Foursquare now officially at 10-million-users. http://techcrunch.com/2011/06/20/foursquare-now-officially-at-10-million-users/.

[26] Techcrunch. Klout Adds Foursquare, But How Much Will It Boost My Score? http://techcrunch.com/2011/08/04/klout-adds-foursquare-but-how-much-will-it-boost-my-score/.

[27] The Economist. Where Networking Works. http://www.economist.com/node/21523437.

[28] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. Design and Evaluation of a Real-Time URL Spam Filtering Service. In *IEEE S&P'11*, 2011.

[29] G. Topa. Social Interactions, Local Spillovers and Unemployment. *Review of Economic Studies*, 68(2):261–95, 2001.

[30] Weka Machine Learning Project. Weka. http://www.cs.waikato.ac.nz/˜ml/weka.